Big Data for Specific Emotion Detection Model: Any Position and Representative Emotion Hashtag Approach

Sanghyub, John Lee (sanghyub.lee@auckland.ac.nz)
The University of Auckland https://orcid.org/0000-0001-6714-0225

JongYoon Lim
University of Auckland

Leo Paas
University of Auckland

Ho Seok Ahn
University of Auckland

Research Article

Keywords: big data, specific emotion, emotion hashtags, emotion labelled tweets, natural language processing

Posted Date: March 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-984542/v1

Abstract

Background/ introduction: We propose the large emotion-labelled dataset consisted of tweets labelling representative emotions hashtags posted over 12 years to train a specific emotion detection model. The dataset is available at https://github.com/ EmotionDetection/6H-AP_emotion_labelled_tweets. Prediction of human emotion has been and remains a major challenge in many research fields such as psychology, neuroscience, and computer science. Tweets are considered as a suitable source for collecting big data using emotion hashtags as reliable emotion annotations. However, little is known about data collection criteria on how to apply emotion hashtags (i.e., type and position of emotion hashtags).

Methods: To elucidate unclear criteria, this paper collected over five million tweets that were divided into six datasets. Five traditional ML algorithms trained on six different datasets were evaluated on both internal test sets (30 analyses) of six datasets and external test set (30 analyses).

Results: We propose the emotion labelled dataset (n =1,478,116; any position of representative emotions hashtags) that achieved the highest F1 score. Furthermore, this paper compared the model trained on the proposed dataset with the model trained on a small dataset. We find that this large dataset further improved the model performance in deep learning (18 analyses) than in traditional ML algorithms (30 analyses). **Conclusions:** Finally, we share the proposed dataset with other researchers to contribute to future specific emotion detection model studies, provide reliable baseline results for this data set.

Introduction

Natural language processing (NLP) is a technology that allows computers to understand human language. Text analysis is a vital technique since most online data is stored as text (e.g., social media, blogs, personal web pages, and product descriptions). To successfully interpret human user opinions in the field of human–robot interaction (HRI), the speech-to-text in which the opinion is expressed must be understood to develop robotics that can emotionally empathize with human users.

Most prior studies applying NLP techniques have focused on classifying sentiment polarity of the texts; for example, positive, negative or neutral sentiment can be extracted from tweets [1–4]. However, human emotions are not merely divided into positive and negative [5]. For instance, assume a publisher attempts to use sentiment analysis for summarising opinions regarding new books. Traditional methods classify both sadness and fear as negative emotions. However, if the readers expressed either of these emotions in a manner where their distinction is significant, they ought to be separated as desirable outcomes. Fear may for example be a desirable emotion when reading a thriller, which may not apply for sadness.

The significance of specific emotion analysis is gaining attention in academia and in commercial sectors. For example, customers' behaviour and emotional states can be predicted from various text sources (e.g., tweets and online reviews), and accurate analysis of textual information may be relevant for various business purposes [6–7]. However, specific emotion pre- diction has been and remains a major challenge because of a high level of subjectivity and limited input sources. This is a greater challenge when only a speech-to-text source is provided without contextual information, such as facial expressions or tone of voice [8]. To overcome these challenges, ML algo- rithms are required to develop and train a specific emotion detection model using textual information.

Several prior studies suggest applying tweets including specific emotion hashtags as datasets to train specific emotion detection models [9–11]. The majority of tweets describe daily life events, expressed via individual post and hashtags. Emotion hashtags are accurate and reliable as these represent the emotions which the writers have directly annotated to their tweets. Contrary, other data sources may require manual annotations or present insufficient labelling. Wang et al. [11] point out that the traditional method of manually labelling emotions in tweets could be inaccurate since the annotators must infer the writers' feelings from the text, thus distorting the writer's original intent. Furthermore, manual annotation of emotions on over a million dataset is virtually impossible. Thus, automatically annotated big data is essential for training ML models that outperform ML models trained on small datasets.

Previous datasets for specific emotion detection model training [11–14] were collected based on different criteria; representative emotion hashtags (e.g., #joy) [13] or synonymous emotion hashtags (e.g., #joy, #jouyful, and #enjoy) [11–12, 14]. In the datasets mentioned above, only the tweets that end with emotion hashtags were selected, yet they failed to demonstrate how filtering hashtags at other positions affect the model performance. Up to now, far too little attention has been paid in investigating the selection criteria for tweets collection.

Andrew [15] suggests that data-centric AI (i.e., focusing on quality data) is more critical than model-centric AI (i.e., focusing on effective ML models (algorithms)) in the development of ML models. Hartung [16] point out that there is a golden rule of "trash in and trash out", meaning that good data is imperative to training a good model. This paper investigates the potency of carefully selected high-quality big data, not only for traditional ML algorithms, but also for latest ML algorithms such as deep learning. The performance of all ML models used in this paper will be cross validated with the internal test set as well as the external test set. Taken together, we address the following four questions:

1. What data collecting criteria (representative or synonymous emotion hashtags) will effectively improve the performance of ML models?
2. Can filtering data based on the position (any, last quarter, or last position) of emotion hashtags in a tweet effectively improve the performance of ML models?
3. Can the large emotion-labelled dataset improve the performance of traditional ML algorithms over a small dataset?
4. What are the differences in the benefits of increasing training data when applying deep learning algorithms compared to traditional ML algorithms?

To answer the above questions, we collected over five million tweets (*n*=5,645,139) by applying 24 synonymous emotion hashtags. The resulting dataset consists of a total of 565,575,630 characters, rendering many novels and books minuscule; e.g., Alice's Adventures in Wonderland is approximately four thousand times smaller (*characters*=142,557). Most prior research on specific emotion detection focused on Ekman's basic emotion theory [13, 17–22]. 24 synonymous emotion hashtags (e.g., #joy, #enjoy, #fun, and #joyful) also belong to six basic emotions [23]; 'fear', 'anger', 'sadness', 'joy', 'surprise' and 'disgust'. Finally, the contributions of this paper as fruitful results of the above questions are as follows.

- Based on rigorous 108 analyzes, we recommend any position of represen- tative emotions hashtags as a clear criterion for an effective emotion-labelled

dataset. This is the surprising finding contrary to previous studies based on last position of synonymous emotion hashtags [11, 12, 14].

- To the best of our knowledge, the proposed emotion-labelled dataset is the largest data set of tweets labelling representative emotions hashtags (*n*=1,478,116) posted over 12 years from OCT 2008 to DEC 2020. We find that this large dataset further improved the model performance in deep learning than in traditional ML algorithms.
- The proposed emotion-labelled dataset is shared with other researchers as a form of open dataset (https://github.com/EmotionDetection/6H-AP_emotion_labelled_tweets, accessed on 16 AUG 2021) to contribute to future specific emotion detection model studies.

The structure of this paper is as follows; Section 2 presents the previous datasets used in specific emotion analysis. Section 3 demonstrates the charac- teristics and exploratory data analysis of collected big data. Section 4 presents pre-processing of big data, introduction of ML algorithms, input data prepara- tion, and evaluation criteria. Section 5 investigates five traditional ML models trained on six datasets with internal and external test sets to propose an effective emotion-labelled dataset. The proposed dataset was then cross eval- uated with a small dataset to show the differences in the benefits of increasing training data between traditional ML and deep learning algorithms. Section 6 concludes this paper.

Related work

The datasets for specific emotion detection models rely on texts to be labelled for specific emotions, but such datasets are virtually impossible to annotate in relevant scale manually. For instance, Liew and Turtle [24] describe that 18 annotators had worked over ten months on labelling 28 emotions in 5,553 tweets. Based on this information, the same annotators would need to work approximately 15 years to manually annotate the dataset collected for the current paper.

There are publicly available datasets, such as Twitter Emotion Corpus (TEC; 21,047 tweets) [13], International Survey on Emotion Antecedents and Reactions (ISEAR; 7,666 sentences) [25], and Affective Text (1,200 news headlines) [26] dataset. Among these datasets, only TEC provides tweets based on representative emotion hashtags (i.e. #fear, #anger, #sadness, #joy, #surprise, and #disgust) at the last position. However, the number of data points is only approximately from 1,200 to 21,047. Such quantity of data is insufficient for ML model development requiring big data [27].

Other researchers collected tweets, including emotion hashtags as an au- tomated emotion annotation. They applied various data collection and data cleaning criteria to collect a large amount of data. For instance, Wang et al. [11] collected a large emotion-labelled (i.e., joy, sadness, anger, love, fear, thankfulness, and surprise) dataset of approximately 2.5 million tweets obtained by searching 131 synonymous emotion hashtags at the last position. The F_1 score was .6163 with large-scale linear classification that classify the seven emotion labels (joy, sadness, anger, love, fear, thankfulness, and surprise). They fur- ther reported that increasing the training data improved the accuracy by .2216 comparing with the small dataset (1,000 tweets).

Saravia et al. [14] collected 664,462 tweets that express eight emotions (sadness, joy, fear, anger, surprise, trust, disgust, and anticipation) with

339 synonymous emotion hashtags at the last position. Their multi-layer convolutional neural network (CNN) architecture with a matrix form of the enriched patterns was trained and achieved an F_1 score of .79.

Abdul-Mageed and Ungar [12] collected 1,608,233 tweets, including 665 synonymous emotion hashtags at the last position across the 24 emotions. Their 'joy' emotion consisted of emotion-related hashtags such as 'happy', 'happiness', 'joy', 'joyful', 'joyfully', and 'delighted'. Their gated recurrent neural nets (GRNNs) model was trained and achieved an accuracy of .8012 based on six basic emotions [23].

However, there has been no detailed investigation into decreasing sample size by filtering data based on the position (any, last quarter, or last position)

of emotion hashtags affect model performance. Also, previous research used different criteria (i.e., representative or synonymous emotion hashtags) for labelling the emotions in the tweets. It is not clear which criteria is optimal. To the best of our knowledge, the four questions we raised earlier are largely unexplored.

Methods

Collecting big data

Anglophone tweets were collected using a Twitter application programming interface (Twitter API) [28]. The application to use the Twitter API has been approved by Twitter for academic research purposes of this paper. Over six million tweets (*n*=6,795,462) posted from MAR 2007 to JUN 2021 were collected by searching six basic emotions [23] since the first tweet was posted on MAR/21/2007 [29].

Table 2 demonstrates six basic emotions consisting of six representative (e.g., #joy) [13] and 18 synonymous emotion hashtags (e.g., #fun, #joyful, and #enjoy) [11, 12, 14]. This paper adopted representative emotion hashtag words (see Table 1) suggested by Mohammad [13], and synonymous emotion hashtag words (e.g., #worried, #pissed, and #eww) suggested by Saravia et al. [14].

Table 1
Tweet samples

Tweets after basic pre-processing	**Hashtag**
Seriously You act like you love me then ignore me Oh yea man that s how we do it	#anger
It makes me sick how many school lock downs have had to happen in the past few days all the press about CT just eggs on copycats	#disgust
I can take a lotta scary stuff but spiders cross the line	#fear
Our family would like to wish you and your beautiful families a very happy prosperous holidayseason	#joy
Sometimes I wish stopping was easier But addiction is just too strong problems addiction selfdepreciative help pls	#sadness
How could a laptop under a blanket even be comfortable cat	#surprise

Table 2

The frequency statistics and length of tweets

Emotions	**Representative**	**Synonymous emotion hashtags**	**Frequency**	**Percent**	**Length**
Anger	#anger	#angry #mad #pissed	1,159,456	20.54	84.96
Disgust	#disgust	#awful #disgusted #eww	780,674	13.83	79.17
Fear	#fear	#feared #fearful #worried	514,452	9.11	88.07
Joy	#joy	#enjoy #fun #joyful	1,102,663	19.53	91.07
Sadness	#sadness	#depressed #grief #sad	1,200,969	21.27	90.16
Surprise	#surprise	#strange #surprised #surprising	886,925	15.71	79.48
Total	6 hashtags	18 hashtags	5,645,139	100.00	85.49

To clear the dataset, emotion hashtag, website address, and special char- acters were removed from tweets as the basic pre-processing (see Table 3) to obtain only English words. Then, all duplicate tweets were removed to keep every tweet unique [12]. Also, tweets that had fewer than three English words and re-tweets were excluded [13]. This resulted in over five million tweets (*n*=5,645,139).

Table 3
Pre-processing process

Process	Pre-processing	Tweet text
Raw	None	that terrible moment when you end an essay with "so... yeah" and then forget to change and ediT IT BEFORE YOU HAND IT IN #anger
Basic	Only English words	that terrible moment when you end an essay with so yeah and then forget to change and ediT IT BEFORE YOU HAND IT IN
Moderate	Lowercased English words	that terrible moment when you end an essay with so yeah and then forget to change and edit it before you hand it in
Rigorous	Stop-words removed	terrible moment end essay yeah forget change edit hand

The TEC dataset [13] consisted of 21,047 tweets posted from NOV 2011 to DEC 2011. There were 558 duplicate cases (2.65%) within the TEC dataset and 731 duplicate cases (3.47%) between our dataset and the TEC dataset after the basic pre-processing. Since the proportion of duplicate cases is small, duplicate cases were not excluded to keep the original dataset.

The collected dataset consisted of publicly available information and did not store any personally identifiable information. This dataset offered two key information (i.e., tweet text as X values and six emotion hashtags as y values; see Table 1). Exploratory data analysis continues in the following section.

Exploratory data analysis

Understanding a dataset plays a vital role in improving model performance and finding missing, incorrect, or biased data. Therefore, unique characteristics and the distributions of the dataset should be identified.

Table 2 shows the frequency statistics and length of tweets (i.e., number of characters of tweets). The most frequent emotion was ‘sadness’ (21%), followed by ‘joy’ (20%) and ‘anger’ (20%), whereas ‘fear’ (9%) was the least prevalent. Each emotion consisted of a representative emotion (e.g., #anger) hashtag and additional three synonymous emotion hashtags (e.g., #angry, #mad, and #pissed).

Two additional points that can be extracted from tweets are: (i) number of characters of a tweet (*Mean* = 85.88, *SD* = 51.42, *Median* = 77, *Min* = 5, *Max* = 336, *Q1* = 49, *Q3* = 107), (ii) number of words of a tweet (*Mean* = 15.41, *SD* = 8.84, *Median* = 14, *Min* = 3, *Max* = 95, *Q1* = 9, *Q3* = 20). Table 2 reports that most tweets were less than the 140 characters limit (280 characters limit from 2017), and ‘joy’ (*Length*=91.07) had the greatest number of characters, whereas ‘disgust’ (*Length*=79.17) had the least number of characters.

[Table 2 about here.]

Figure 1 shows the word cloud of the most frequent and occasionally words after processing the rigorous pre-processing (see Table 3). Interestingly, this figure shows that ‘love’ (*n*=328,589) was one of the most mentioned emotion words, which implies positive valence, but do not reflect to any specific emotion.

Few tweets included multiple synonymous emotion hashtags (e.g., I am #anger #fear; 2.06%; *n*=116,398), and most tweets did not include multiple different emotion hashtags (e.g., I am #anger #joy; .11%; *n*=6,422). Also, one of ten tweets included synonymous emotion words (e.g., I am anger #fear; 13.05%; *n*=737,550), whereas most tweets did not include different emotion words (e.g., I am anger #joy; .30%; *n*=17,233).

In summary, these results show that predicting emotions with emotion words or other frequent words, such as 'now' (*n*=233,949), 'don' (*n*=217,213), and 'one' (*n*=213,669), might be difficult. These results also suggest applying deep learning algorithms that can focus on the contextual meaning of a sentence rather than traditional ML algorithms that can focus on words.

[Fig. 1 about here.]

Dataset preparation process

This section introduces the preparation of the emotion-labelled dataset (*n* = 5,626,219) collected in the previous section. The dataset preparation process consists of three steps: 1) pre-processing, 2) dataset selection, and 3) dataset splitting strategy. The purpose of pre-processing is to uniformly organise the input text to enhance the recognition, performance, and efficiency of the model. Two pre-processing processes (the moderate and the rigorous process) were applied to the text datasets. As shown in Table 3, in the moderate process, English words were lowercased after processing the basic process. This process was applied to deep learning models that recognise the contextual meaning of the entire sentence. Although, stop-words are frequently used words without special meaning, such as 'we', 'are', 'the', 'a', 'only', and 'in', were not removed as they can play a significant role in conveying meaning for deep learning models. In the rigorous process, stop-words were removed after processing the moderate process. This process is applied to traditional ML models that focus on words rather than a sentence.

[Table 3 about here.]

Secondly, the pre-processed dataset was divided into six datasets according to dataset selection criteria (i.e., type (representative or synonymous emo- tion hashtags) and position (any, last quarter, or last position) of emotion hashtags). For examples of position of emotion hashtags; 'I am john #joy' represents last position, 'I am #joy john' represents last quarter, and '#joy I am john' represents any position. Table 4 provides detailed information about the six different datasets of that abbreviations mean; 24H: 24 hashtags, 6H: 6 hashtags, AP: any position, LQ: last quarter, and LP: last position.

Table 4
Six different datasets

Id	Type	Hashtags	Position	N
24H-AP	synonymous	24	AP	5,645,139
24H-LQ	synonymous	24	LQ	4,023,748
24H-LP	synonymous	24	LP	2,183,452
6H-AP	representative	6	AP	1,478,116
6H-LQ	representative	6	LQ	903,002
6H-LP	representative	6	LP	390,630

Finally, the 80/20 dataset splitting strategy for model training and validation was applied to the six datasets. This split ratio is commonly used for NLP-related or other ML tasks [1]. For example, the 24H-AP dataset (*n*=5,645,139) was split into 80% for training (*n*=4,516,111) and 20% for testing (*n*=1,129,028).

To evaluate the generalizability of the models' performance, the TEC dataset was applied as an external test dataset. As noted in the related work section, TEC dataset provides emotion-labelled tweets with representative emotion hashtags at the last position. TEC dataset (*n*=21,047) was also split into 80% for training (*n*=16,837) and 20% for testing (*n*=4,210). Also, traditional and deep learning ML algorithms trained on the six different datasets including the TEC dataset were evaluated.

Traditional machine learning algorithms

Five ML algorithms (i.e., 1) k-nearest neighbours, 2) decision tree, 3) naive bayes, 4) support vector machine, and 5) logistic regression) [30] trained on six

different datasets were applied to propose the effective dataset. Below these algorithms are briefly introduced.

First, the K-Nearest Neighbours (KNN) algorithm is the most widely used centroid-based clustering algorithm. It is an unsupervised learning technique that automatically groups data with similar characteristics into respective clusters. This algorithm is called K-NN because it generates k individual clusters, and the output is the value of the object's feature, and the centre points of k clusters represent the average value of the shortest distance from the data in the cluster [31].

Second, the Decision Tree (DT) algorithm is widely used in data science due to its various advantages such as excellent predictive accuracy, intuitive description of a model, and selecting informative attributes in model design. The analysis result of DT can be drawn as a tree diagram that groups properties by sorting the various data entries by information gain (informative attributes). DT is primarily used for classification purposes [31].

Third, the Naive Bayes (NB) algorithm based on the Bayes rule is mainly used for clustering and classification purposes. The underlying architecture is based on conditional probability. Depending on the likelihood of happening, trees are created, which are also called Bayesian networks [31]. NB is mainly used for text classification, such as sentiment analysis [32–34].

Fourth, the Support Vector Machine (SVM) algorithm is one of the most widely used ML algorithms. SVM is mainly used for classification. SVM works according to the principle of margin calculation and basically draws as much margin as possible between the data to be classified [31]. The purpose of the SVM algorithm is to find the hyperplane in an N-dimensional space that clearly classifies data points [30].

Fifth, the Logistic Regression (Logit) algorithm uses a simple algorithm (sigmoid functions) like ordinary least squares (OLS). The relationship be- tween dependent and independent variables is expressed as a mathematical function and used in future prediction models. The main difference from OLS is that the results are divided into specific categories [35]. However, this single-

layer perceptron may not be suitable for dealing with high dimensional prob- lems of big data with many observations and an indefinite number of variables.

Deep Learning algorithms

Three representative deep learning algorithms (i.e., 1) artificial neural network,

2) recurrent neural network, and 3) convolution neural network) [36] were applied to further evaluate the proposed effective dataset.

First, the Artificial Neural Network (ANN) with multi-layer perceptron (MLP) technology has become the most advanced ML technology available today. It has been particularly successful in areas such as voice recognition, image analysis, and natural language processing [37]. ANN is divided into three layers (i.e., input layer, hidden layer, and output layer). A hidden layer improves the accuracy of predictions by enabling the classification of complex structures that may be found in big data.

Second, the Recurrent Neural Network (RNN) algorithm can find patterns in the sequences of sentence and classifies the results when receiving sequence data such as a sentence [38]. It extracts a specific pattern for a sentence by sequentially inputting text information without using the given word feature vector such as the traditional ML algorithms and ANN discussed above. In the RNN, the previously input information is gradually accumulated in the hidden state and transmitted to the current input state, thereby enabling predictive modelling of sequence data.

Third, the CNN algorithm can stack multiple convolutional layers. In general, it is used to classify images by learning to extract the best features by applying various filters to the input image [39]. Yoon [40] demonstrates that CNN can be applied not only to image data but also to text data with outstanding classification performance. While RNN reflects the input order of words in training, CNN classifies sentences by reflecting the appearance information of words in each sentence to training.

Count vectorising to extract features from text

Since words cannot be provided as input data and only numerical data may be used to train ML models, it is necessary to convert words or sentences into specific numeric values through feature extraction. This paper adopts the count vectorizer; a method of constructing a word vector after measuring the number of times a word or words appear.

Specifically, n-gram (n consecutive words) [36] was set to 3-gram where up to 3 consecutive words were included in the word vector. For example, when providing two texts as input data, such as ’John is happy’ and ’John is angry’, feature names (i.e., ’angry’, ’happy’, ’is’, ’is angry’, ’is happy’, ’john’, ’john is’, ’john is angry’, ’john is happy’) are assigned. Accordingly, the array value of ’John is happy’ is assigned as (0, 1, 1, 0, 1, 1, 1, 0, 1). Six different datasets and TEC dataset were count vectorised to train ML models.

F1 score for multi-class classification

The purpose of ML models trained with the word vector is to perform multi- class classification in which the output indicates the likelihood of the input sentence being classified as one of the six emotion labels (i.e., ‘fear’, ‘anger’, ‘sadness’, ‘joy’, ‘surprise’ and ‘disgust’). ML models were evaluated by using the F_1 score. Although classification accuracy is widely applied due to its easy measurement, it has been criticised for being unsuitable for application to real-world problems, as its simplicity disables measurement of imbalanced data [30]. Vinodhini and Chandrasekaran [41] point out that the F_1 score is suitable measure of the ML models tested with imbalanced data, which calculates the harmonic balance between precision and recall taken from the confusion metric. The formula for the F_1 score of each label (class) can be expressed as:

$$\mathrm{F_1score}(l) = \frac{2 \times \mathrm{precision}(l) \times \mathrm{recall}(l)}{\mathrm{precision}(l) + \mathrm{recall}(l)}$$

, where *l* is the label (i.e., anger, disgust, fear, joy, sadness, and surprise), precision(*l*) is calculated as $\frac{\mathrm{truepositive}\,(l)}{\mathrm{truepositive}\,(l) + \mathrm{falsepositive}\,(l)}$, and recall(*l*) is calculated as $\frac{\mathrm{truepositive}\,(l)}{\mathrm{truepositive}\,(l) + \mathrm{falsenegative}\,(l)}$.

The F1 scores calculated for each label were weighted according to the number of data points in each label to derive the weighted F_1 score [42], applying to the proposed model evaluation. Overall, a total of eight ML algorithms (i.e., five traditional ML and three deep learning algorithms) were evaluated using the highest weighted F_1 score (see Fig. 2).

[Fig. 2 about here.]

Results

The purpose of the following sections is to evaluate five traditional ML algorithms trained on the six different datasets, both on internal (section 5.1) and external test sets (section 5.2). The purpose of this analysis is to propose an effective emotion-labelled dataset. Then, five traditional ML (section 5.3) and three deep learning algorithms (section 5.4) trained on the proposed and TEC train sets were evaluated, both on proposed and TEC test sets. The purpose of this analysis is to show that the proposed large dataset further improved the model performance in deep learning than in traditional ML algorithms.

Internal evaluation by traditional ML algorithms on six test sets

The purpose of this section is to evaluate the traditional ML algorithms (Logit, SVM, NB, DT, and KNN) trained on six datasets (24H-AP, 24H-LQ, 24H- LP, 6H-AP, 6H-LQ, and 6H-LP), which results in a total of 30 analyses. The rigorous pre-processing and the 80/20 dataset splitting strategy were applied. The input variable was word vectors of count vectorised tweets with 3-gram, and max-features was set to 100,000 where low-frequency words after the

100,000th were excluded. The output variable concerned the six emotion labels (i.e., ‘fear’, ‘anger’, ‘sadness’, ‘joy’, ‘surprise’ and ‘disgust’).

Traditional ML algorithms were utilised using scikit-learn [43]; one of the most widely used ML libraries in the Python programming language. All models proposed in this paper were trained on a computer with NVIDIA RTX 3090 GPU with 24GB memory.

[Fig. 3 about here.]

Figure 3 shows the average F_1 scores of six datasets on internal test sets. What stands out in this figure is that increasing synonymous emotion hashtags and decreasing sample size by filtering the hashtag location (LQ and LP) decrease the classification accuracy of the internal test set. Detailed F_1 scores of the five traditional ML algorithms (see Table 5) also show the same patterns (6 hashtags > 24 hashtags and AP > LQ > LP). These results suggest that 6H-AP dataset (any position of representative emotion hashtags) could be an effective emotion-labelled dataset.

Table 5
F_1 scores of 30 ML models on internal test sets

Dataset	6H-AP	6H-LQ	6H-LP	24H-AP	24H-LQ	24H-LP
Logit	.6783	.6531	.5721	.5762	.5618	.5165
SVM	.6739	.6390	.5418	.5863	.5684	.5133
NB	.6679	.6322	.5153	.5687	.5460	.4860
DT	.5762	.5438	.4507	.4771	.4510	.4011
KNN	.4958	.4498	.3889	.4136	.3843	.3536
Total	**.6184**	**.5836**	**.4938**	**.5244**	**.5023**	**.4541**

External evaluation by traditional ML algorithms on TEC

The 30 ML models trained in the previous section were externally evaluated with an external data set, TEC, to assess the generalizability of the models' performance. The rigorous pre-processing and 80/20 dataset splitting strategy were applied to TEC dataset. Accordingly, an external test set of 4210 tweets was applied for the 30 analyses.

[Fig. 4 about here.]

Figure 4 presents the average F_1 scores of six datasets on the external test set, TEC. As can be seen, the pattern (6 hashtags > 24 hashtags and AP > LQ > LP) was replicated as when evaluated with the internal test set. Detailed F_1 scores of the 30 ML models on the external test set (see Table-6) also show

the same patterns under equal circumstances. ML models trained on 6H-AP dataset consisting of any position (AP) of emotion hashtags outperformed the TEC test set consisting of the last position (LP) of emotion hashtags.

Table 6
F_1 scores of 30 ML models on external test set, TEC

Dataset	**6H-AP**	**6H-LQ**	**6H-LP**	**24H-AP**	**24H-LQ**	**24H-LP**
Logit	.5665	.5686	.5582	.4693	.4644	.4545
SVM	.5660	.5601	.5313	.4915	.4811	.4519
NB	.5924	.5922	.5359	.4844	.4504	.3399
DT	.5269	.5109	.4882	.4651	.4367	.3988
KNN	.4450	.4437	.4009	.3802	.3633	.3151
Total	**.5394**	**.5351**	**.5029**	**.4581**	**.4392**	**.3920**

Cross-evaluation by traditional ML algorithms on 6H-AP and TEC

This section aims to demonstrate that the traditional ML models trained on the proposed large dataset (6H-AP) can improve F_1 scores rather than the traditional ML models trained on small dataset (TEC). As shown at the top of Fig. 2, five cross-evaluations were performed using five traditional ML algorithms trained on two datasets (6H-AP and TEC). The rigorous pre-processing and 80/20 dataset splitting strategy were applied. The input variable was the word vectors of count vectorised tweets with 3-gram, and max-features was set to 100,000. The output variables were the six emotion labels.

[Fig. 5 about here.]

The differences in model performance between models trained on 6H-AP and models trained on TEC are highlighted in Table 7. It is found that models trained on the large dataset (6H-AP) outperformed their smaller dataset counterpart (TEC). Fig. 5 provides the average F1 scores of cross-evaluations by traditional ML algorithms. The results in Table 7 show that models trained on 6H-AP achieved a higher average weighted F_1 score of .2545 (= 6H-AP on 6H-AP (.6184) – TEC on 6H-AP (.3639)) than the models trained on TEC. The results on TEC shows that models trained on 6H-AP achieved a higher average weighted F_1 score of .0453 (= 6H-AP on TEC (.5394) – TEC on TEC (.4941)) than models trained on TEC. Finally, the small train (1%6H-AP) set was obtained by randomly selecting 16,837 cases from the 6H-AP train set (*proportion*=1.42%). The results in Table 7 show that models trained on the large dataset (6H-AP) also outperformed models trained on the small dataset (1%6H-AP). The results in this section indicate that increasing training data brings a solid performance improvement of the ML models.

Table 7 F_1 scores of 15 traditional ML models

Train set	**6H-AP**	**6H-AP**	**TEC**	**TEC**	**1%6H-AP**	**1%6H-AP**
Test set	6H-AP	TEC	6H-AP	TEC	6H-AP	TEC
Logit	.6783	.5665	.4351	.5605	.5742	.4571
SVM	.6739	.5660	.4526	.5652	.5475	.4370
NB	.6679	.5924	.3032	.4840	.5676	.4426
DT	.5762	.5269	.4010	.4915	.4931	.3787
KNN	.4958	.4450	.2277	.3692	.3014	.2480
Total	**.6184**	**.5394**	**.3639**	**.4941**	**.4968**	**.3927**

Cross-evaluation by deep learning algorithms on 6H-AP and TEC

The purpose of this section is to demonstrate that the proposed large dataset further improved the model performance in deep learning (e.g., ANN, CNN, and RNN) than in traditional ML algorithms. Three cross-evaluations were performed using three deep learning algorithms trained on two datasets (6H-AP and TEC) to compare with the traditional ML models in previous section. The moderate pre-processing and 80/20 dataset splitting strategy were applied. The input variable was word vectors of using tokenized tweets, and max-features was set to 100,000. The output variable was six emotion labels. Deep learning models were utilised using Tensorflow [44]; one of the most widely used deep learning libraries in the Python programming language.

In the structure of ANN model, the input layer was connected to the embedding layer for word vectors. The max-pooling layer was then used, followed by two fully connected hidden layers attached to the output layer. The structure of CNN model was modified from the structure recommended by Yoon [40]. The input layer was connected to the embedding layer for word vectors. A total of three convolutional layers were then used, and values were extracted with different filter sizes (e.g., 3, 4, and 5) followed by the max-pooling layer. This model was completed by stacking up a dropout layer to prevent overfitting and two fully connected hidden layers connected to the output layer (i.e., six emotion labels). The structure of RNN model was modified from the structure recommended by Kumar et al. [36]. The input layer was connected to the embedding layer for word vectors. A total of two bidirectional long short-term memory (Bidirectional LSTM) [45] layers were then used. This model was completed by stacking up a dropout layer to prevent overfitting and two fully connected hidden layers connected to the output layer.

When training three models with 6H-AP, the batch size was set to 128, the learning rate was set to .00001, and the dropout rate was set to .2. To prevent overfitting, the training of the models was finished in approximately 16 to 30 epochs. When training other models with TEC and 1%6H-AP, the batch size was set to 4, the learning rate was set to .0001, and the dropout rate was set to .2. To prevent overfitting, the training of the models was finished in approximately 4 to 5 epochs.

[Fig. 6 about here.]

Figure 6 illustrates the average F1 scores of cross-evaluations by deep learning algorithms. This figure grants a multitude of insights. First, the pattern (6H- AP on 6H-AP > 6H-AP on TEC > TEC on TEC > TEC on 6H-AP) was

replicated as when evaluated by traditional ML algorithms. Second, models trained on 6H-AP showed similar or slightly better performance improvement than the Logit algorithm that showed superior performance among traditional ML algorithms (see Table 7 and Table 8). Third, models trained on TEC showed similar or slightly lower performance degradation as compared to Logit and SVM algorithms (see Table 7 and Table 8). Finally, models trained on the large dataset (6H-AP) also outperformed models trained on the small dataset (1%6H-AP).

Table 8
F_1 scores of nine deep learning models

Train set	**6H-AP**	**6H-AP**	**TEC**	**TEC**	**1%6H-AP**	**1%6H-AP**
Test set	6H-AP	TEC	6H-AP	TEC	6H-AP	TEC
ANN	.6791	.5620	.3925	.5165	.5361	.4369
CNN	.7013	.5788	.4720	.5580	.5538	.4258
RNN	.6942	.5773	.4415	.5455	.5559	.4199
Total	**.6915**	**.5727**	**.4353**	**.5400**	**.5486**	**.4275**

Conclusions

This paper proposes the application of big data (emotion-labelled tweets) for training specific emotion detection models. Previously, little attention has been paid in investigating the selection criteria for tweets collection. Five traditional ML models trained on six datasets were evaluated, on both internal and external test sets. We found that the 6H-AP dataset (any position of representative emotions hashtags) is an effective emotion-labelled dataset. This finding is contrary to previous studies based on synonymous emotion hashtags [11–12, 14] and last position of emotion hashtags [11–14]. Finally, we cross-evaluated the 6H-AP dataset and the small datasets (TEC and 1%6H- AP).

This paper shows that this large dataset can make a greater contribution to improving model performance in deep learning than in traditional ML algorithms. This new understanding should help to improve the performance of the specific emotion detection model by using the proposed 6H-AP dataset and by using deep learning algorithms that can train large and complex data sets. The 6H-AP dataset is, to the best of our knowledge, the largest emotion-labelled dataset available with any position of representative emotions hashtags applied.

This paper poses some limitations. We tested a tweet dataset with specific emotion hashtags annotated. However, researchers need to be aware of the shortcomings of directly deriving from the nature of online user-generated content (UGC). Several biases are described in the literature that analyses UGC dataset, such as self-selection and response biases [46]. Also, the specific emotion hashtags annotated by writers may reflect unusual or special circum- stances. Nevertheless, with the proposed large sample dataset, this paper can assume that it represents the entire population rather than the outliers within a small sample, allowing the proposed dataset to train an unbiased and robust model.

The question raised by this study is whether predicting human emotions differ from artificial intelligence. For example, human emotion prediction using surveys may provide deeper insights into the differences and characteristics of artificial intelligence and human emotion prediction. Another further study could assess whether the performance of the deep learning models can be improved when the state-of-the-art pre-trained transformer language models are applied. This research has emerged many inquiries relevant in our current paradigm and needs further investigation.

Declarations

Acknowledgements

The authors would like to thank the Twitter for providing developer API through an academic partnership, and the CARES team and the University of Auckland Robotics Department for their support.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

This work was not funded.

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Availability of data and material

The proposed emotion-labelled dataset is shared with other researchers as a form of open dataset (https://github.com/EmotionDetection/6H- AP_emotion_labelled_tweets, accessed on 16 June 2021) to contribute to future specific emotion detection model studies.

Code availability

Not applicable

Authors' contributions

SJ Lee drafted the manuscript and designed the study.

All authors considered the results and approved the final manuscript.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable.

References

1. Lim J, Sa I, Ahn HS, Gasteiger N, Lee SJ, MacDonald B. Subsentence Extraction from Text Using Coverage-Based Deep Learning Language Models. Sensors. 2021;21(8):2712–2712. Available from: https://dx.doi. org/10.3390/s21082712.
2. Wang B, Liakata M, Zubiaga A, Procter R. TDParse : Multi-target-specifi sentiment recognition on Proceedings of the 15th Conference of the European Chapter. 2017;1.
3. Dong L, Wei F, Tan C, Tang D, Zhou M, Xu Adaptive recursive neural network for target-dependent twitter sentiment classification. Acl-2014. 2014;p. 49–54.
4. Vo DT, Zhang Y. Target-dependent twitter sentiment classification with rich automatic IJCAI International Joint Conference on Artificial Intelligence. 2015;p. 2015–2015.

5. Ahn HS, Choi JY. Can we teach what emotions a robot should express. 2012 IEEE/RSJ International Conference on Intelligent Robots and Sys- tems. 2012;p. 1407–1412.
6. Berger J, Milkman KL. What Makes Online Content Viral? Journal of Marketing Research. 2012;49(2):192–205. Available from: https://dx.doi. org/10.1509/jmr.10.0353.
7. Rocklage MD, Fazio The Enhancing Versus Backfiring Effects of Positive Emotion in Consumer Reviews. Journal of Marketing Re- search. 2020;57(2):332–352. Available from: https://dx.doi.org/10.1177/ 0022243719892594.
8. Chatterjee A, Narahari KN, Joshi M, Agrawal SemEval-2019 task 3: EmoContext contextual emotion detection in text. Proceedings of the 13th International Workshop on Semantic Evaluation. 2019;p. 39–48.
9. Hasan M, Rundensteiner E, Agu Emotex: Detecting emotions in twitter messages. 2014 ASE Bigdata/Socialcom/Cybersecurity Conference. 2014;.
10. Mohammad SM, Kiritchenko Using Hashtags to Capture Fine Emotion Categories from Tweets. Computational Intelligence. 2015;31(2):301–326. Available from: https://dx.doi.org/10.1111/coin.12024.
11. Wang W, Chen L, Thirunarayan K, Sheth Harnessing twitter" big data" for automatic emotion identification. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. 2012;p. 587–592.
12. Abdul-Mageed M, Ungar L. Emonet: Fine-grained emotion detection with gated recurrent neural networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017;1:718–728.
13. Mohammad S. # Emotional tweets. * SEM 2012: The First Joint Conference on Lexical and Computational 2012;1:246–255.
14. Saravia E, Liu HCT, Huang YH, Wu J, Chen YS. Carer: Contextual- ized affect representations for emotion recognition. Proceedings of the 2018 Conference on Empirical Methods in Natural Language 2018;p. 3687–3697.
15. Andrew Deeplearning AI issue 84; 2021. Available from: https://www. deeplearning.ai/the-batch/issue-84/.
16. Hartung Making big sense from big data in toxicology by read-across. ALTEX-Alternatives to animal experimentation. 2016;33:83–93.
17. Hajar M. Using YouTube comments for text-based emotion recognition. Procedia Computer Science. 2016;83:292–299.
18. Jain VK, Kumar S, Fernandes SL. Extraction of emotions from multilin- gual text using intelligent text processing and computational linguistics. Journal of Computational Science. 2017;21:316–326.
19. Li X, Pang J, Mo B, Rao Y. Hybrid neural networks for social emotion detection over short text. 2016 International Joint Conference on Neural Networks (IJCNN). 2016;p. 537–544.

20. Luyckx K, Vaassen F, Peersman C, Daelemans W. Fine-Grained Emotion Detection in Suicide Notes: A Thresholding Approach to Multi-Label Clas- sification. Biomedical Informatics 2012;5s1:BII.S8966–BII.S8966. Available from: https://dx.doi.org/10.4137/bii.s8966.

21. Roberts K, Roach MA, Johnson J, Guthrie J, Harabagiu SM. EmpaTweet: Annotating and Detecting Emotions on Lrec. 2012;12:3806–3813.

22. Sen A, Sinha M, Mannarswamy S, Roy Multi-task representation learning for enhanced emotion categorisation in short text. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2017. p. 324–336.

23. Ekman Universals and Cultural Differences in Facial Expression of Emotions, Nebasaka. In: Symposium on Motivation. University Nebaska Press; 1972. p. 83–207.

24. Liew JSY, Turtle Exploring fine-grained emotion detection in tweets. Proceedings of the NAACL Student Research Workshop. 2016;p. 73–80.

25. Scherer KR, Wallbott Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology. 1994;66(2):310–328. Available from: https://dx.doi. org/10.1037/0022-3514.66.2.310.

26. Strapparava C, Mihalcea R. Learning to identify emotions in text. Proceedings of the 2008 ACM symposium on Applied computing. 2008;p. 1556–1560.

27. Siegel E. Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons; 2013. .

28. Kishore S, Peko G, Sundaram D. Looking Through the Twitter Glass: Bridging the Data-Researcher Gap. AMCIS 2019 Proceedings. 2019;.

29. Search api | twitter api | docs | twitter developer platform; 2021. Available from: https://developer.twitter.com/en/docs/twitter-api/ enterprise/search-api/overview.

30. Provost F, Fawcett Data Science for Business: What you need to know about data mining and data-analytic thinking. Reilly Media, Inc; 2013. .

31. Dey A. Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies. 2016;7(3):1174–1179.

32. Kang H, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. 2012;39(5):6000–6010. Available from: https://dx.doi. org/10.1016/j.eswa.2011.11.107.

33. Ye Q, Zhang Z, Law Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications. 2009;36(3):6527–6535.

34. Zhang Z, Ye Q, Zhang Z, Li Y. Sentiment classification of Internet restau- rant reviews written in Cantonese. Expert Systems with Applications. 2011;38(6):7674–7682.

35. Tu Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology. 1996;49(11):1225–1231.

36. Kumar N, Dangeti P, Bhavsar K. Natural language processing with Python cookbook. Packt Publishing; 2019. .
37. Zhang Q, Yang LT, Chen Z, Li A survey on deep learning for big data. Information Fusion. 2018;42:146–157.
38. Leeflang PS, Wieringa JE, Bijmolt TH. Advanced methods for modeling markets. Berlin: Springer; 2017. .
39. Zhang Y, Wallace A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:151003820. 2015;.
40. Yoon Convolutional neural networks for sentence classification. arXiv preprint arXiv:14085882. 2014;.
41. Vinodhini G, Chandrasekaran RM. Sentiment analysis and opinion min- ing: a International Journal. 2012;2(6):282–292.
42. Chakravarthi BR, Priyadharshini R, Muralidaran V, Suryawanshi S, Jose N, Sherly E, et al. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. Forum for Information Retrieval 2020;p. 21–24.
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of Machine Learning Research. 2011;12:2825–2830.
44. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016;.
45. Graves A, Schmidhuber J. Framewise phoneme classification with bidi- rectional LSTM and other neural network architectures. Neural Net- works. 2005;18(5-6):602–610. Available from: https://dx.doi.org/10.1016/ j.neunet.2005.06.042.
46. Stamolampros P, Korfiatis N, Kourouthanassis P, Symitsi E. Flying to Quality: Cultural Influences on Online Reviews. Journal of Travel Re- search. 2019;58(3):496–511. Available from: https://dx.doi.org/10.1177/ 0047287518764345.

Figures

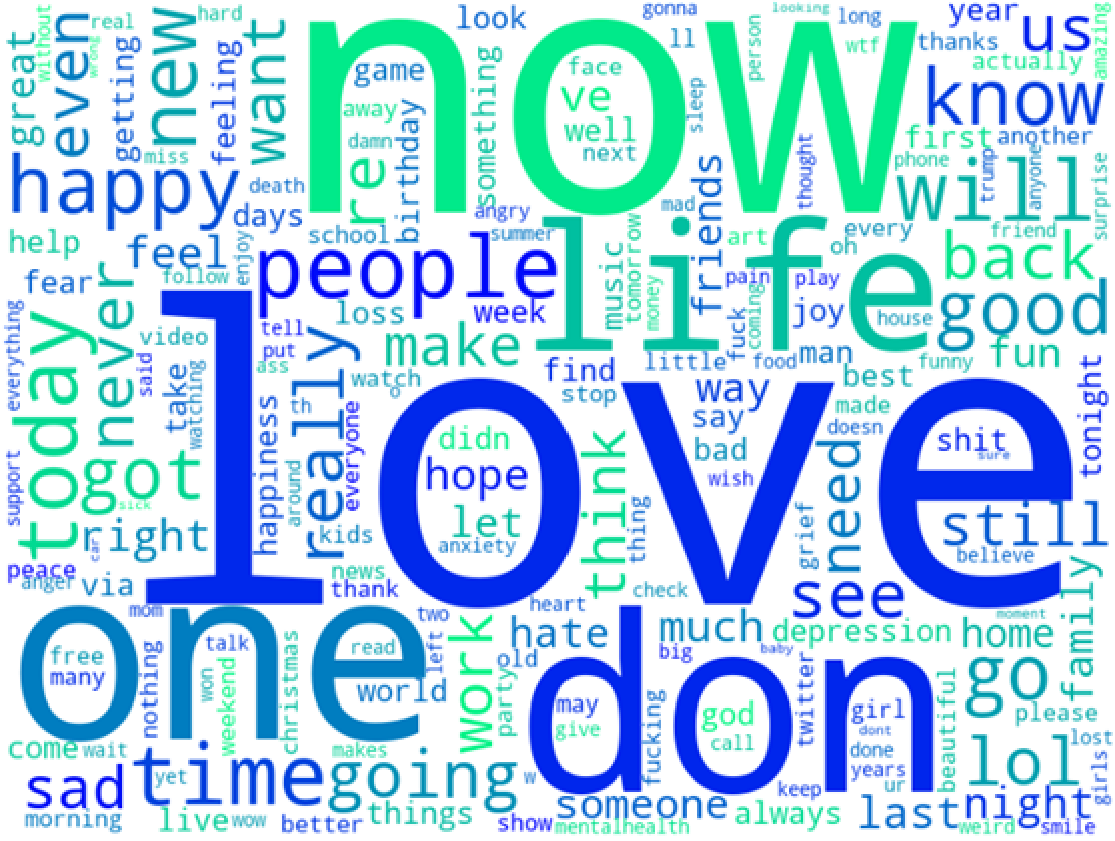

Figure 1

Word cloud of the most frequent and occasionally words

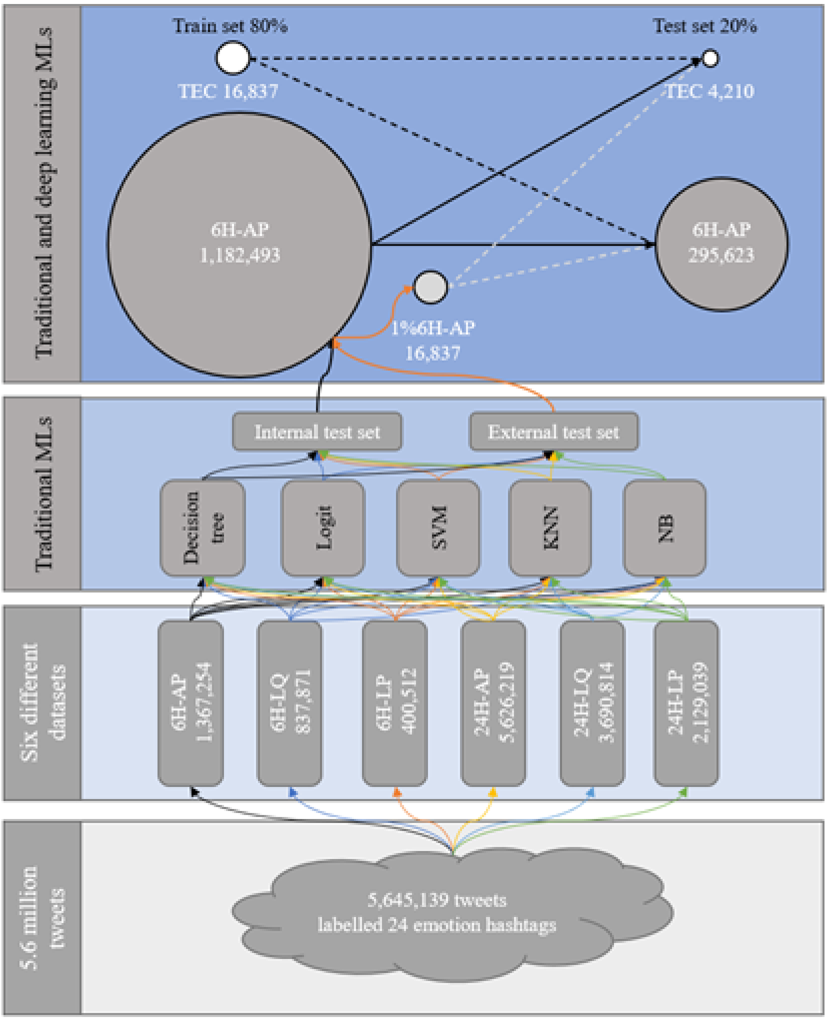

Figure 2

The overview of the important steps used to evaluate the emotion-labelled datasets:

First, 5,645,139 tweets labelled 24 emotion hashtags were divided into six datasets. Second, five traditional ML algorithms trained on six different datasets were evaluated on internal and external test sets to propose an effective emotion-labelled dataset, namely 6H-AP dataset. Finally, five traditional ML algorithms trained on the large and small train sets were evaluated on proposed and TEC test sets. Then, three deep learning algorithms trained on the large and the small sets were evaluated on proposed and TEC test sets. The results show that the proposed large dataset further improved the model performance in deep learning than in traditional ML algorithms.

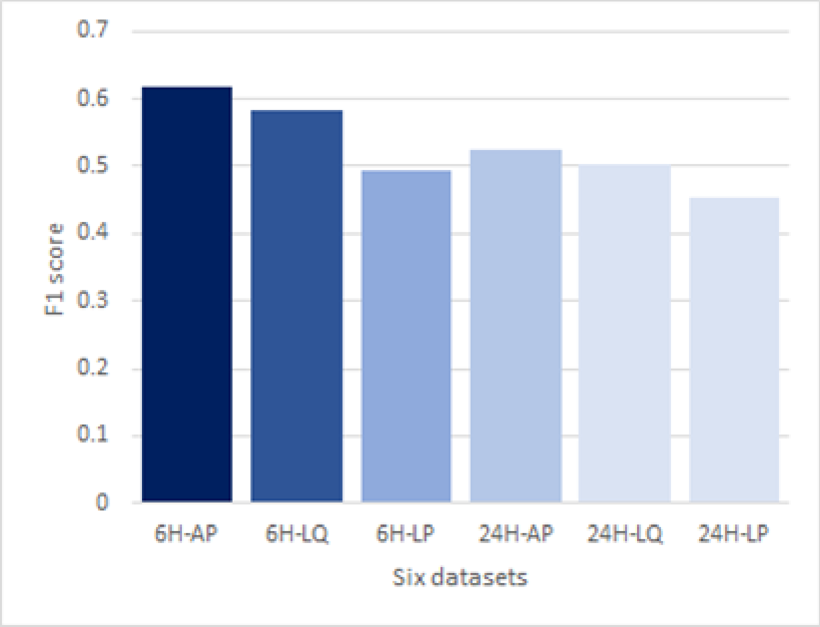

Figure 3

Average F_1 scores of six datasets on internal test sets

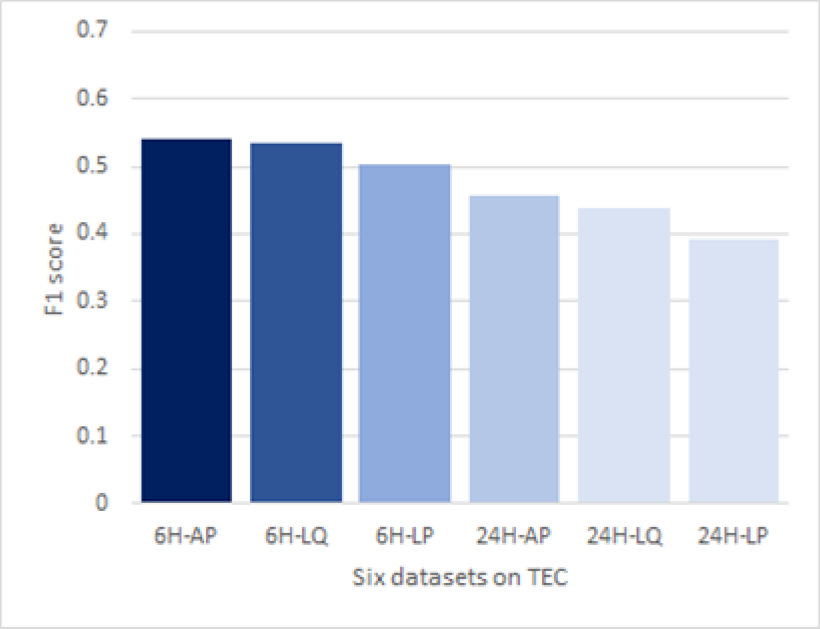

Figure 4

Average F_1 scores of six datasets on the external test set, TEC

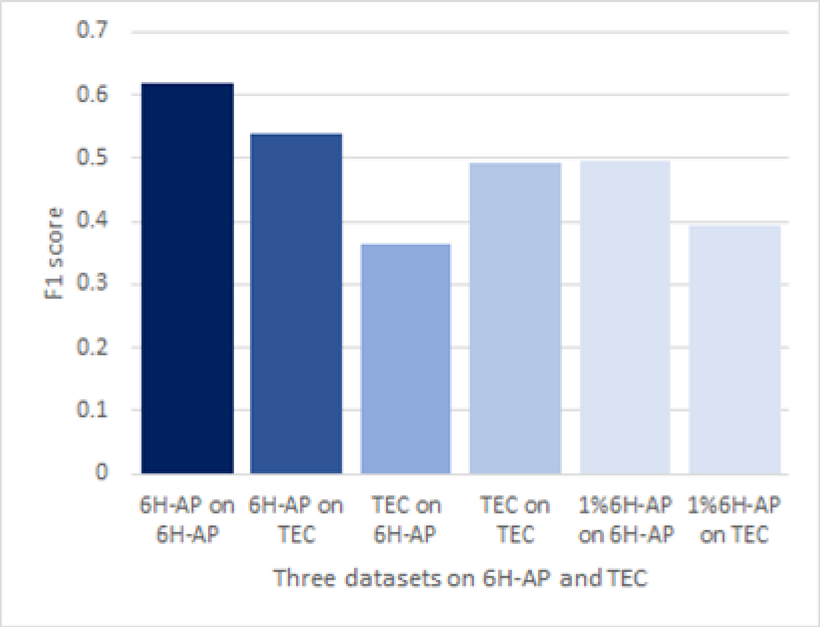

Figure 5

Average F_1 scores of cross-evaluations by traditional ML algorithms

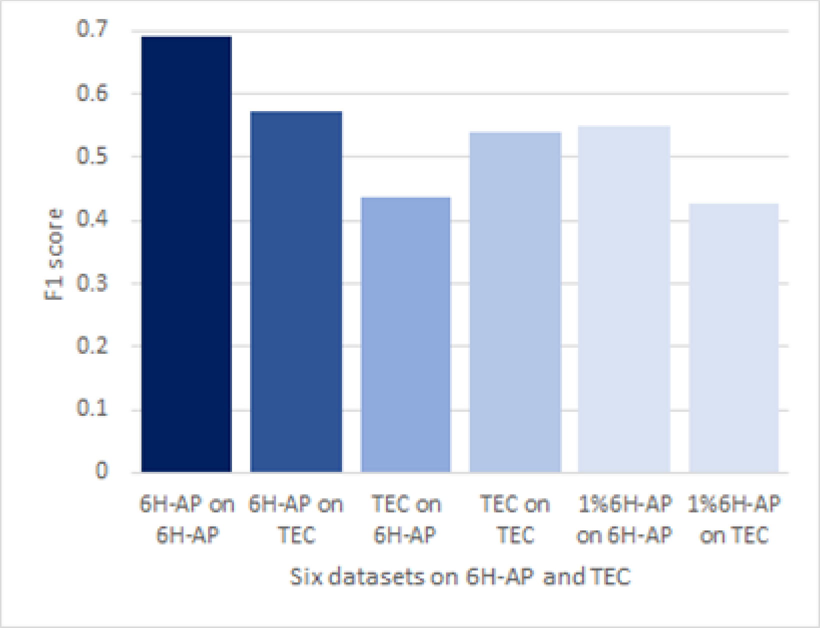

Figure 6

Average F_1 scores of cross-evaluations by deep learning algorithms